Survey of Information Retrieval Techniques for Web using NLP

Rini John
Department of Computer Engineering,
Mumbai University
PIIT, New Panvel, India

Sharvari S. Govilkar
Department of Computer Engineering,
Mumbai University
PIIT, New Panvel, India

ABSTRACT

Web is loaded with information and is getting overloaded with data each passing day. There needs to be efficient development in the area of information retrieval so the required data can be fetched accurately and efficiently. In this paper the two promising areas Natural Language Processing and Web technologies which can be combined together to enable the enterprise to combine the unstructured and structured data in ways that was not handled efficiently by traditional tools. The better understanding of web information can be done by integrating NLP and Web portal. The paper also explores various techniques used both areas. Also the NLP frameworks which can be used for the future work in this area.

Keywords

NLP, information Retrieval, semantic Assistance, entity extraction, visual page segmentation, semi-markov conditional random fields, hierarchical conditional random field and Web Portal.

1. INTRODUCTION

Natural Language Processing is an area concerned with the interface between human natural languages and computer. It's part of Computer Science, Artificial Intelligence and Computational Linguistics. Summarization, Sentiment analysis, Auto-categorization, search are many areas of the branches of Natural Language Processing. Entity extraction has gain importance in recent times due to information overload from tons of Webpages. This is happening because single entity information can be queried and the data about this entity can be obtained through thousands of Webpages. To improve people's browsing experience it is extremely important to understand the structure and semantics of Webpage. This paper explores the various developments in the field of information retrieval in Web using NLP.

Information retrieval is a process of getting the desired data accurately and efficiently. However the question is how to combine NLP and several semantic technologies to help users in creating knowledge, analyzing and renewing bulk amount of contents. To get quantifiable improvements for the users, the open question is the way to integrate system like web portals, webpages with the upcoming technologies.

The paper presents a survey of various information retrieval techniques, NLP frameworks and its application. Information retrieval is defined in section II. The past literature has been citied in section III. Information retrieval techniques for Web using NLP are discussed in section IV. Summary of above Techniques are given along with application in section V. The conclusion of the paper is given in section VI.

2. INFORMATION RETRIEVAL

Information retrieval is a process through which user gets the desired information from a pool of data resources. Based on full-text indexing or metadata the searches are done. When a user queries a system for the relevant data that is the start point of information retrieval process. It has gained much more importance due to massive and increasing resources in the Web. In an era of maximum use of internet, information retrieval has gained utmost importance.

3. LITERATURE SURVEY

In this section the significant past literature that use different information retrieval technique of Web using NLP. Most of the researchers have combined techniques of the fields to get the most effective results. Techniques are need which would focus more on the semantic portions in a web page. In the previous work in these areas, tag-tree is represented by tag structure mainly used to denote a web page. Instead of concentrating on the content structure more attention is given in the presentation structure.

Deng Cai, Shipeng Yu, Ji-Rong Wen, and Wei-Ying Ma has proposed Vision-based Web Entity Extraction[1] using VIPS Algorithm (a Vision-based Page Segmentation Algorithm). Tag trees tend to focus on the presentation structure instead of the content structure which is the main issues with previous works as they often are not correct enough to differentiate in the web page the semantic portions. Also designers have different style to compose a web page which is intricate and varied. This paper proposes Vision-based page segmentation (VIPS Vision-based page segmentation (VIPS) approach to overcome these problems.

It does the page segmentation based on human perception and uses various page layout features like font-size, different colors used in the sections of a webpage to build a vision-tree for a page.

Jun Zhu, Zaiqing Ni, Ji-Rong We, Bo Zhang, Wei-Ying Ma has introduced a Hierarchical Conditional Random Field (HCRF)[2] model for understanding a page layout. To get efficient and accurate results on information retrieval of entities the importance of Page-layout understanding is an absolute necessity. With Vision-tree, nodes are the resultant output but assigning the labels becomes a task. It includes the long distance dependencies to achieve promising results.

Identification of entity is one of the most important feature information retrieval. To get required information for the specified query can only be obtained if the entities are well defined. William W. Cohen has introduced Semi-CRF [3] in which according to the assigned labels the text content inside the html element is segmented to identify the entities much more finer and accurate way. Also the output is comprehensive portrayal about the entities as whole together. In the higher-order models the computational cost is high which immensely reduced in Semi-CRF and it gives much of the outcomes of these models. It is an extension of CRFs. Instead of measuring the properties of individual elements it measures property of segments.

To integrate the common desktop applications, mobile applications, NLP, web information systems, a Semantic Assistants project [5] has been developed which focuses directly bringing services to the end users. To enable this integration of GATE NLP services with client, a service-oriented architecture is developed. This framework can be taken into consideration for the future developments for information retrieval through web to gain advantage of the various NLP techniques.

Paolo Nesi, Gianni Pantaleo and Marco Tenti [9] have presented Geographic information Extraction from web documents and text data. They have developed GeLo systems which help organizations and companies to extract geographical coordinates and addresses from their web domains. The values of Recall measures, Precision and F-measure show promising results. Thus these encouraging outcomes can be taken into account for the several problems faced in geographical information retrieval.

The authors Suma Adindla and Udo Kruschwitz [10] have combined NLP and IR for intranet search. They have encompassed a dialogue component displaying the usefulness of a task based assessment in search system. The main target area is local web sites. The authors has proposed a system that directs the user in the search process by getting the data collection and mining the pieces of knowledge from the documents which is automatically getting populated in database.

Zhong Liu and Ying Wang[11] have developed a novel method of Chinese Web Information Extraction and Applications for retrieving semantically rich information from the unstructured and semi-structured Chinese web pages, Zhong and Ying have presented a work flow of their IE system. To extract pattern and built knowledge repository the approach used are Knowledge engineering approach and automatic training.

Ruiqiang Guo and Fuji Ren[12] have analyzed in this paper the link between NLP and Semantic Web. They have explained how NLP benefit Semantic Web to implement information retrieval.

B.Aysha Banu and Dr.M.Chitra[13] have proposed a novel ensemble vision based deep web data extraction technique for Web Mining Applications. In this paper issues regarding information retrieval of the contents from the web pages have been highlighted. The prevailing methods have certain boundaries dealing with huge number of web pages and also the programming is language (HTML) dependent. To solve this problem Vision based approach is used to retrieve information from the web pages. It assists the user instinctively to partition the webpage into number of semantic parts where the visual and spatial features are vital to this process.It emphasis on the primary visual features of a web page.

I.Vijayalakshmi and Sobha Lalitha[14] have provided a search facility for documents containing text which are in mobile or web and how to retrieve them is presented here. For the implementation they have considered English and Tamil documents for information retrieval.

4. INFORMATION RETRIEVAL TECHIQUES FOR WEB USING NLP

Entities are extracted relevant to the domain to get entities like people, location, and organization. Following are various information retrieval Techniques and frameworks used for web using NLP.

4.1 VIPS (Visual Page Segmentation)

Information retrieval can take advantage from this page structure as VIPS uses tag-tree free method to get the content structure. There are three components of the VIPS algorithm:

4.1.1 Visual Block Extraction

From the current subpage, the goal is to find the appropriate visual blocks. As per the intra visual difference of every extracted node, the DoC value is fixed. Until all the proper nodes are found, this process is reiterated to get the visual blocks of the present sub-page.

4.1.2 Visual Block Extraction

Visual separator detection algorithm is run to identify the separator between the block and relation among them; this is done to get the separated block by running this algorithm for each block.

4.1.3 Content Structure Construction

Content structures can be constructed when the weights are established and detected of the separators. The lowest weight separators of the blocks are the criteria to create new blocks by initiating the construction process.

4.2 Hierarchical conditional random field (HCRF)

In Hierarchical Condition Random Field (HCRF) helps in labeling that is identifying and assigning labels to the HTML elements. Following example can be considered for better understanding; full name and other address details are given of a particular entity.

PUTHENPARAMBIL RINI JOHN
4 - Viraj Vihar,
Road no.48
VADODARA GUJARAT (GJ) – 410555

Then using HCRF the outcome of the first line that is full name will be

MIDDLENAME_FIRSTNAME_LASTNAME
➔ PUTHENPARAMBIL_RINI_JOHN

The output of the HCRF model is graph of webpage where junction tree algorithm is used to understand labeling on the graph of vertices. The graph vertices are the nodes on the vision tree.

4.3 Semi-CRF (Semi-Markov conditional random fields)

Semi-CRF is used for segmentation of an input sequence and assigning labels to these segments. Following example can be considered to understand more clearly, a Full name is given which includes first name, middle name and last name.

PUTHENPARAMBIL RINI JOHN

Then using Semi-CRF the following outcomes

MIDDLENAME_FIRSTNAME_LASTNAME
➔ MIDDLENAME → PUTHENPARAMBIL
➔ FIRSTNAME → RINI
➔ LASTNAME → JOHN

It is an addition to the linear chain CRF in which iterative process of labeling is done for segments.

4.4 Semantic Assistants Framework

NLP services are applied in GATE framework. It captures the artifacts, their languages and tasks.Following are the various tiers of Semantic Assistants:

Tier 1: Client-Side Abstraction Layer (CSAL) and client applications.

Tier 2: NLP Service Connector and Web server.

Tier 3: GATE API is used to get the request from the client, language services assemble. This is maintained by GATE framework.

Tier 4: Web Ontology Language (OWL) which consist of language services. This tier has external documents which is accessed by the NLP system.

4.5 Web Portals and NLP

A combined access to several information resources and facilities is provided by a web portal. It's a web application for example like My Yahoo! where user finds variety of services like news, entertainment, maps and numerous other information. Commonly used portal technology Java Portlet Specification JSR2862 in which various Application Programming Interface. This API is used for creating portlet applications. This infrastructure consists of server, container and portlets.

A server is intermediary component which is functioning between container and the client. Its responsible for yielding the user request to the from the client to container. The container provides the environment for execution and the life cycle of instances are handled here. Also communication between portlets and preferences are stored in the container.

4.6 GeLo

The main aim of developing this structure is to retrieve information, mining and geolocalizing the web domains related to various research institutes and companies. This model works jointly with NLP techniques to achieve the desired results. The modular architecture of GeLo system [9] is consist of three main blocks

Distributed Web crawler for mining and fetching huge amounts of documents and Big Textual Data.

NLP based linguistic parser and Pattern matching technique is responsible for digging out information associated with the organization owing to the web domain like any geographic information, addresses or location.

Addresses or geographical data are retrieved from Semantic smart city repository.

4.7 Dialogue-Driven Intranet Search

Suma Adindla and Udo Kruschwitz have combine NLP and IR for Intranet Search [10] for creating interaction between user and university intranet search engine just with ease as the user do. Here a dialogue component is fused into search system which is the one of the main component.

There are two parts into this search system: offline knowledge extraction for extracting named entities and predicate argument structures. For mapping the query into the mined knowledge, online mapping process is used.

Document collection is automatically is used to build a domain model where it is processed with help of NLP tools such as Annie IE system which is the component of Gate NLP which is also used for identifying entities and information extraction techniques.Stanford parser is used for extracting simple facts. The knowledge base against which all these elements are brought together to get the user query. The dialogue manager which is one of the vital component in charge of the online mapping process.

4.8 Information extraction system for Chinese Web information extraction

A novel method of Chinese Web Information Extraction and Applications [11] has been developed to mine pattern and identify various entities and their relationships. Amalgamation of various technologies has been used by Zhong Liu for self-training knowledge systems and language segmentation.

The appropriate facts which needs to extracted from the web pages from the Chinese web and Pattern extraction algorithm to get the related correct information to build the pattern repository. The pattern for the semi-structure text is retrieved through approach of Knowledge engineering.

Entities recognition task is done by using method maximum forward boundary recognition method. To segment word the approach used is ICTCLAS(Institute of Computing Technology, Chinese Lexical Analysis system)

4.9 Towards the Relationship between Semantic Web and NLP

Ruiquiang Guo and Fuji Ren[12] have discussed how NLP results can be improved with help of Semantic Web technologies in Information extraction. A collection of RDF graphs can be viewed as Semantic Web. To improve the results in IE, optimized query strategies and effective saving of the data are the areas which need an immediate attention.

Different query methods have been discussed to effectively query the concerned data. The query language XQUERY, SQL and SPARQL are most common query language used to get the effective results. Numerous information extraction which are based on ontology tools like Controlled language are used. It helps the users to get the status reports and interpret meeting minutes.

4.10 A Novel Ensemble Vision Based Deep Web Data Extraction

B.Aysha Banu, Dr.M.Chitra[13] have developed a approach to extract the content structure of web which is built on visual representation. This result obtained through this approach is helpful for many applications of information retrieval.

To retrieve data from deep Web pages to get structured outcomes for multi data-region deep web pages they have used an approach vision based. The following strategy is used to get the desired results, first a Visual Block tree is obtained by using Visual page segmentation algorithm for a given web page which makes the usage of layout features of the page and creates a division of the page at a semantic level. From the original page which consist of logical content structures through which nodes are extracted. Appearance, layout and position features are used upon which data records, locations are based.

4.11 Automatic Information Extraction Through Mobile

The authors I.Vijayalakshmi and Sobha Lalitha Devi[14] for creating the IE engine, a rule base approach has been used here. They have given the main focus on Web data and SMS data for this IE engine. To get the quality data in end, the Information Extraction algorithm is one of main aspect concentrated to get significant information from the documents.

All the keywords which are specific to the domain are initialized first along with the rules related to each domain specific keywords. For each of the sentence the file is read and the next step is assembling domain specific keywords and saving it. Later both the database rules are checked for the rules in sentences. Matching rules are extracted from fields and the others are not selected. This process is iterated for the other sentences

Table 1. Summary of WebNLP Techniques

Sr.no.	WebNLP Techniques	Description	Application
1	Visual Layout Features	Deng Cai and the co-authors have implemented VIPS Algorithm (a Vision-based Page Segmentation Algorithm) which creates a vision –tree of the web page which is based on the content –structure rather than the presentation structure.	Search Engine to get the relevant information based on the concept of the human perception.
2	Hierarchical conditional random field (HCRF)	Jun Zhu and the co-authors has introduced a model called HCRF-Hierarchical conditional random field model which labels the vision tree nodes of html elements. Thus helps is identification of the elements.	Data mining, to generate relations between various entities and extract useful data. In search engines so that the entities are labeled semantically.
3	Semi-CRF (semi-Markov conditional random fields)	The author William W. Cohen has introduced Semi-CRF which can be used in further segmentation of the nodes of the elements to get finer and accurate results.	Information Retrieval applications where accuracy is very important like the weather forecast, search of criminal records based on the relations etc.
4	Semantic Assistants Framework[8]	This paper discusses about the various NLP services which are incorporated in GATE framework.	Framework can be used to integrate with the various applications like Desktop application to get the various features of NLP. Also can used with various application text summarization, extraction.
5	Web Portals and NLP	This paper shows how Semantic assistant framework can be helpful in information retrieval in Web portals.	Various porlets can take the advantage of this integration between the NLP Enabled Content Portlets.
6	GeLo	This framework has been designed with the help of various NLP techniques the goal of mining, retrieving and geolocalizing web domains associated to company's and research institutes.	The evaluation of such system shows high values for F-Measure, Precision and Recall measures. This showing promising results for taking into account the problem faced in retrieving geographical information retrieval from unstructured data.
7	Dialogue-Driven Intranet Search	Dialogue component in a search system is used for assessing the usefulness of task-based evaluation. They have particularly targeted local web sites such university intranets. The idea applied is a dialogue system extracts the small pieces knowledge from the document collection that can then be mapped against the query.	The application of these kind of system can be used in Intranet search where the importance is given to largely automated knowledge extraction process using NLP.
8	Information extraction system for Chinese Web information	Here the design used is object-level vertical search system for IE where	They have developed a object level vertical search system for

	extraction	the results demonstrated that Information extraction is effective improved through this search system. It focus on optimizing knowledge repository and attempting to build people relationship matrix.	Chinese people search system.

5. CONCLUSION

This paper summarizes the techniques which will be helpful in retrieving the desired data from the web. The Web is getting a massive overload of information each day. So these techniques and framework provide the new direction to improve entity extraction. We believe the techniques surveyed in this paper can be combined due to the promising results showed. Also various frameworks are discussed which can be used for various needs like geographical extraction, Intranet search and Portals. VIPS helps in understanding a page in a way human perceive which is the important point to be considered as in the near future these types of techniques would be helpful to get accurate results also the promising outcomes as compared to previous algorithm are commendable.

For the further development in future the VIPS techniques along with HCRF and Semi-CRF should be considered to get effective results as the output of VIPS is a tree structure which can be used by HCRF for identification and labeling of the entities. Also in the end Semi-CRF can be applied to the results of HCRF to get further segmentation to get more precise and effectual results. Also further improvement can be achieved with the help of parallel processing to get the results in desired time frame. The grouping of Portal and Web jointly along with various NLP techniques can benefit users to find desired and appropriate content and quickly grip the main points.

6. ACKNOWLEDGMENTS

It is a great pleasure and moment of immense satisfaction for me to express my profound gratitude to Prof. Sharvari Govilkar whose constant encouragement enabled me to work enthusiastically. Her perpetual motivation, patience and excellent expertise in discussion during progress of the survey work have benefited me to an extent, which is beyond expression.

7. REFERENCES

[1] D. Cai, S. Yu, J.-R. Wen, and W.-Y. Ma, “VIPS: A vision-based page segmentation algorithm”, Microsoft Tech. Rep., MSR-TR-2003-79, 2003.

[2] J. Zhu, Z. Nie, J.-R. Wen, B. Zhang, and W.-Y. Ma, “Simultaneous record detection and attribute labeling in web data extraction”, in Proc. Int. Conf. Knowl. Disc. Data Mining, 2006, pp. 494–503.

[3] S. Sarawagi and W. W. Cohen, “Semi-Markov conditional random fields for information extraction”, in Proc. Conf. Neural Inf. Process. Syst., 2004, pp. 1185–1192.

[4] Fedor Bakalov, Bahar Sateli, Ren´e Witte, Marie-Jean Meurs, Birgitta K, “Natural Language Processing for Semantic Assistance in Web Portals”, 2012

[5] R. Witte and T. Gitzinger, “Semantic Assistants – User-Centric Natural Language Processing Services for Desktop Clients,” in 3rd Asian Semantic Web Conference (ASWC 2008), ser. LNCS, vol. 5367. Bangkok, Thailand: Springer, 2008, pp. 360–374. [Online].

[6] http://en.wikipedia.org/wiki/Natural_language_processing

[7] http://en.wikipedia.org/wiki/General_Architecture_for_Text_Engineering

[8] http://www.semanticsoftware.info/semantic-assistants-architecture

[9] Paolo Nesi,Gianni Pantaleo and Marco Tenti, “Ge(o)Lo(cator):Geographic Information Extraction from Unstructured Text Data and Web Documents”, in 2014 9th International Workshop on Semantic and Social Media Adaption and Personalization.

[10] Suma Adindla and Udo Kruschwitz, “Combining the Best of Two Worlds: NLP and IR for Intranet Search”, in 2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology.

[11] Zhong Liu and Ying Wang, “A Novel method of Chinese Web Information Extraction and Applications”, in 2009 WASE International Conference on Information Engineering.

[12] Ruiqiang Guo and Fuji Ren, “Towards the Relationship Between Semantic Web and NLP”,2009

[13] B.Aysha Banu, Dr.M.Chitra , “A Novel Ensemble Vision Based Deep Web Data Extraction” in 2012 IEEE Intemational Conference on Advanced Communication Control and Computing Technologies (ICACCCT)

[14] I.Vijayalakshmi, Sobha Lalitha Devi, “Automatic Information Extraction through Mobile” ,in ICCCNT'12 26th_28th July 2012, Coimbatore, India